# The assessment of academic style in EAP writing: The case of the rating scale

**Ute Knoch**
**University of Melbourne**

## Abstract:

Rating scales for EAP writing assessment sometimes make reference to academic style or register. Typical descriptors might describe students' writing as 'lacking academic style' or displaying an 'adequate understanding of academic style'. These descriptors do not offer much guidance to raters in the rating process and potentially result in different interpretations and foci on the side of the raters.

A number of researchers (e.g. Crismore, Markkanen, & Steffensen, 1993; Hyland, 1997, 2000b, 2002b, 2002c; Hyland & Milton, 1997) have attempted to measure academic style more objectively. However, these efforts have not thus far been reflected in rating scale descriptors. For the purpose of this study, a variety of discourse analytic measures of reader/writer interaction were used to analyse 602 academic writing scripts at a variety of proficiency levels. Based on the results of this analysis, a new rating scale was formulated. The study investigates whether such an empirically-grounded scale can be used to assess academic style in students' writing more reliably and with greater discrimination than the more traditional measure. The validation process involves a multifaceted Rasch analysis of scores derived from multiple ratings of 100 scripts using the old and new rating descriptors as well as a qualitative analysis of questionnaires canvassed from the raters. The results suggest that raters were able to apply the more detailed descriptors more reliably and consistently. The findings are discussed in terms of their implications for rating scale development and rater training.

*Melbourne Papers in Language Testing* 2008 Volume 13.1 pp. 34-67. The Language Testing Research Centre, The University of Melbourne.

## 1. Introduction:

Because writing assessment requires subjective evaluations of writing quality by raters, the raw score candidates receive might not reflect their actual writing ability. One reason for the variability found in writing performance might lie in the way rating scales are designed. Fulcher (2003) has shown that most existing rating scales are developed based on intuitive methods which means that they are either adapted from already existing scales or they are based purely on what developers think might be common features in the writing samples in question. However, for rating scales to be more valid, it has been contended that rating scales should be based on empirical investigation of actual writing samples (Fulcher, 1987, 1996; North, 2003; North & Schneider, 1998; Turner & Upshur, 2002; Upshur & Turner, 1995, 1999).

## 2. The assessment of academic style in writing

Hyland (2000b) argues that writers do more than produce texts in which they present an external reality; they also negotiate the status of their claims, present their work so that readers are most likely to find it persuasive, and balance fact with evaluation and certainty with caution. Writers have to take a position with respect to their statements and to their audiences, and a variety of features have been examined to see how they contribute to this negotiation of a successful reader-writer relationship. This attempt at creating a reader-writer relationship in EAP contexts is often also referred to as academic style.

Academic style has been given little attention in rating scales for EAP writing assessment. The IELTS rating scale for Task 2 (www.ielts.org), for example, mentions in the task response category that the writing piece might have an inappropriate format. The category for lexical resource mentions inappropriately used vocabulary. However, as is the case with

most other rating scales, academic style is not further mentioned, although anecdotal evidence suggests that examiners at times refer to it.

In the fields of second language acquisition research and discourse analysis, several researchers have attempted to operationalize academic style in a way that it can be analysed more objectively in student writing. Crismore et al. (1993), for example, created a taxonomy of interpersonal metadiscoursal markers. These are divided into the following categories:

1. Hedges (epistemic certainty markers)
2. Certainty markers (epistemic emphatics or boosters)
3. Attributors
4. Attitude markers
5. Commentaries

**Hedges**, have been defined as 'ways in which authors tone down uncertain or potentially risky claims' (Hyland, 2000a), as 'conventions of inexplicitness' and as 'a guarded stance' (Shaw & Liu, 1998), as structures that 'signal a tentative assessment of referential information and convey collegial respect for the views of colleagues' (Hyland, 2000a) or as 'the absence of categorical commitment, the expression of uncertainty, typically realized by lexical devices such as *might'* (Hyland, 2000b). Examples of hedges are epistemic modals like *might, may, could*, and other structures such as *I think, I feel, I suppose, perhaps, maybe, it is possible*. Hyland (2000b) suggests that hedges are highly frequent in academic writing and are more frequent than one in every 50 words.

A number of researchers have looked at hedging in L2 learners' writing. Bloor and Bloor (1991), for example, found that direct and unqualified writing rather than the use of hedging devices, was more typical of EFL writers. Similarly, Hu, Brown and Brown (1982) found that Chinese L2 writers are more direct and authoritative in tone and make more use of stronger modals than native speakers. Hyland and Milton (1997) investigated how both L1 and L2 students express doubt and certainty in

writing. They found that the two groups of writers used a similar number of modifiers - one device in every 55 words - but native speakers used two-thirds of the devices to weaken claims whilst non-native speakers used over half of the modifiers in their writing to strengthen claims. In a more recent study, Kennedy and Thorp (2002) were able to show that writers at levels four and six in the IELTS writing section used fewer hedging devices than writers at level eight.

***Boosters (or certainty marker*s)** have been defined as expressions 'that allow writers to express conviction and to mark involvement and solidarity with an audience' (Hyland, 1998) or as 'the ways in which writers modify the assertions they make, emphasizing what they believe to be correct' (Hyland, 2000a). Boosters include expressions like *clearly show, definite, certain, it is a fact that* or *obviously*. As has already been described above in the context of hedges, a number of studies found that L2 writers overuse boosters in their writing and are therefore found to make unjustifiably strong assertions (Allison, 1995; Bloor & Bloor, 1991; Hyland & Milton, 1997; Kennedy & Thorp, 2002).

The third structure on Crismore's list of interpersonal discourse markers, ***attributors***, increase the force of an argument and can take the form of a narrator as in '*John claims that the earth is flat'* or as an attributor as in '*Einstein claimed that our universe is expanding'*. In Vande Kopple's (1985) categorization, these were separate categories, but Crismore et al. (1993) found in their analysis that these two features performed a very similar function and therefore grouped them together.

***Attitude markers*** express the writer's affective values and emphasize the propositional content, but do not show commitment to it. These include words and phrases like '*unfortunately*' or '*most importantly*'. They can perform the functions of expressing surprise, concession, agreement, disagreement and so on.

Finally, the category of ***commentaries*** establishes a reader-writer relationship by bringing the reader into the discourse through expressions like '*you may not agree that*', '*my friend'*, '*think about it'*.

Intaraprawat and Steffenson (1995) used all the categories described above to investigate differences between good and poor ESL essays. They found that good students used twice as many hedges, attitude markers and attributors, more than double the number of emphatics (boosters) and three times as many commentaries.

Apart from hedges, boosters, attributors, attitude markers and commentaries, writers can also express reader-writer interaction by showing ***writer identity*** in their writing. As Hyland (2002a) suggests, academic writing is not just about conveying an ideational 'content', it is also about the representation of self. Hyland (2002c) argues that L2 writers are often told not to use 'I' or 'in my opinion' in their academic writing. In his investigation on the use of the first person in L1 expert and L2 writing, he found that professional writers are four times more likely to use the first person than L2 student writers (Hyland, 2002a).

Hyland (2002c) argues that this underuse of first person pronouns in L2 writing inevitably results in a loss of voice. Contrary to Hyland's (2002a; 2002c) findings, Shaw and Liu (1998) showed that as L2 students' writing develops, they move away from using personal pronouns in their writing and make more use of passive verbs. They therefore argue that more developed writing has less authorial reference.

If writers choose not to display writer identity, but rather want to keep a piece of writing more impersonal, they could do this by increased use of the ***passive voice***. This was investigated by Banerjee and Franceschina (2006), who found that the higher the IELTS score awarded to a writing script, the more passives the writer had used.

## 3. The study:

The aim of this study is to investigate whether the markers of reader/writer interaction described above can successfully be operationalised into a rating scale for academic style. The study was undertaken in two phases. Firstly, 602 writing samples were analyzed to establish whether markers of reader/writer interaction were able to distinguish between writers at five levels of writing ability. The findings were then transferred into a rating scale. To validate this scale, ten raters were trained and then rated 100 writing samples. The findings were compared to previous ratings of the same 100 scripts by the same raters using an existing rating scale for academic style. After the rating rounds, raters were given a questionnaire to canvass their opinions about the rating scale and a subset of seven raters were interviewed. A visual presentation of the study can be seen in Figure 1 below.

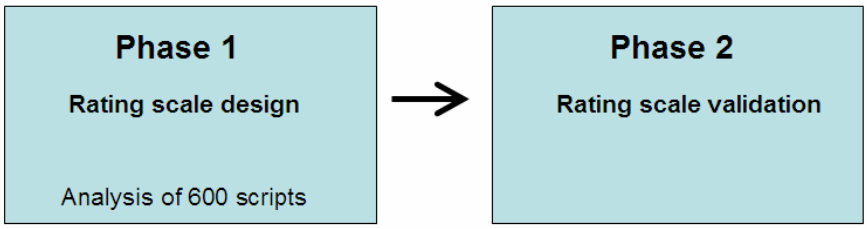


**Figure 1: Outline of the study**

The research questions were as follows:

1) What are the features of reader/writer interaction at different levels of expository writing?
2) How reliable and valid is an empirically-developed rating scale for academic style when compared to a pre-existing, more traditional measure?

3) What are rater's perceptions of the two rating scales?

## 4. Method:

**4. 1. Context of the research**

This study was conducted in the context of DELNA (Diagnostic English Language Needs Assessment) which is administered at the University of Auckland, New Zealand. DELNA is a university-funded procedure designed to identify the English language needs of undergraduate students following their admission to the University, so that the most appropriate language support can be offered. DELNA is administered to both native and non-native speakers of English. This context was selected by the researcher purely because of its availability and because the rating scale used to assess the writing task (see description below) is representative of many other rating scales used in EAP writing assessment across the world. A more detailed description of the assessment and the rating scale can be found in the section below.

*4. 1. 1. The assessment instrument*

The DELNA assessment includes a screening component which consists of a speed-reading and a vocabulary task. This is used to eliminate highly proficient users of English and exempts these from the time consuming and resource-intensive diagnostic procedure. The diagnostic component comprises objectively scored reading and listening tasks and a subjectively scored writing task.

The writing section is an expository writing task in which students are given a table or graph of information which they are asked to describe and interpret. Candidates are given a time limit of 30 minutes to complete the task. The writing task is routinely double (or if necessary triple) marked analytically on nine traits (Organization, Coherence, Academic style, Data description, Interpretation, Development of ideas,

Sentence structure, Grammatical accuracy, Vocabulary & Spelling) on a six point scale ranging from four to nine. The assessment criteria were developed in-house, initially based on an existing scale. A number of validity studies have been conducted on the DELNA assessment battery, which included validation of the rating scale (Elder, 2003; Elder & Erlam, 2001; Elder & von Randow, 2002). The wording of the scale has been changed a number of times based on the feedback of raters after training sessions or during focus groups. The DELNA rating scale reflects common practice in performance assessment in that the descriptors are graded using adverbs like '*adequate*', '*appropriate*', '*sufficient*', '*severe*' or '*slight*'. The rating scale for academic style uses descriptors like '*adequate understanding of academic style*' or '*style not appropriate to task*'.

*4.1.2. The writing samples*

To identify the specific features of reader/writer interaction used by writers taking DELNA, 602 writing samples, which were produced as part of the 2004 administration of the assessment, were randomly selected. The samples were originally hand-written by the candidates. The mean number of words for the scripts was 269, ranging from 75 to 613.

*4.1.3. The candidates*

329 of the writing samples were produced by females and 247 by males (roughly reflecting the gender distribution of DELNA) whilst the remaining 26 students did not report their gender. The L1 of the students (as reported by a self-report questionnaire) was varied. 42% (or 248 students) have an Asian first language, 36% (217) are native speakers of English, 9% (52) are speakers of a European language other than English, 5% (31) have either a Pacific Island language or Maori as first language and 4% (21) speak either an Indian or a language from Sri Lanka as first language. The remaining 4% (22) were grouped as other. Eleven students did not fill in the self-report questionnaire. The scripts used in this analysis were all rated by two DELNA raters. In case of discrepancies

between the scores, the scores were averaged and rounded (in the case of a .5 result after averaging, the score was rounded down). The 602 scripts were awarded the following average marks (Table 1):

**Table 1: Score distribution of 602 writing samples**

| DELNA score | Frequency | Percent |
|---|---|---|
| 4 | 23 | 4% |
| 5 | 115 | 19% |
| 6 | 253 | 46% |
| 7 | 172 | 29% |
| 8 | 26 | 4% |

*4.1.4. The raters*

The eight DELNA raters taking part in this study were drawn from a larger pool of raters based on their availability at the time of the study. All raters have high levels of English proficiency although not all are native speakers of English. Most have experience in other rating contexts, for example, as accredited raters of the International English Language Testing System (IELTS), whereas others have gained rating experience in other contexts. All have post-graduate degrees in either English, Applied Linguistics or Teaching English to Speakers of other Languages (TESOL). All raters have several years of experience as DELNA raters and take part in regular training moderation sessions either in face-to-face sessions or online (Elder, Barkhuizen, Knoch, & von Randow, 2007; Elder, Knoch, Barkhuizen, & von Randow, 2005; Knoch, Read, & von Randow, 2007).

**4.2. Procedures – analysis of writing samples**

An initial pilot study showed that, probably due to the nature of the discourse, very few attributors, attitude markers and commentaries were

used in the essays. These categories were therefore excluded from any further analysis.

Hedges, boosters, markers of writer identity and the use of the passive voice was investigated by using the concordancing program MonoConc (Barlow, 2002). The complete list of items investigated for hedges, boosters and markers of writer identify was established can be found in Appendix 1. Each lexical item was investigated individually using MonoConc. Here special care needed to be taken, so that lexical items that did not function as hedges or boosters were excluded from the analysis. For example, in the case of the booster *certain*, all uses of *certain* + noun needed to be excluded as this structure does not act as a boosting device. In the case of the lexical item *major*, all uses of the word in conjunction with *cities* or *axial routes*, for example, needed to be excluded because these were also not used as boosters. So for each lexical item in Appendix 2, the whole concordancing list produced in MonoConc needed to be thoroughly examined before each instance of that item could be entered into a spreadsheet. Finally, all items were added together, so that a final frequency count for each script was found for hedges, boosters and markers of writer identity. The passive voice was initially also investigated using MonoConc. However, because it was impossible to search for erroneous instances of the passive (i.e. unsuccessful attempts), this analysis was later refined by a manual search.

**4. 3. Procedures – Rating scale validation**

The raters rated one hundred scripts using the current DELNA criteria and then the same one hundred using the new scale. The scripts were selected to represent a range of proficiency levels. They all participated in a rater moderation session to ensure they were thoroughly trained. All raters were further instructed to rate no more than ten scripts in one session to avoid fatigue.

After rating the two sets of one hundred scripts, the raters filled in a questionnaire canvassing their opinions about the scales. The

questionnaire (part of a larger-scale study) allowed the raters to record any opinions or suggestions they had with respect to the new scale. The questionnaire questions were as follows:

1) What did you like about the scale?
2) Were there any descriptors that you found difficult to apply? If yes, please say why.
3) Please write specific comments that you have about the scales below. You could for example write how you used them, any problems that you encountered that you haven't mentioned above or you can mention anything else you consider important.

A subset of seven raters was also interviewed after the study was concluded.

The results of the two rating rounds were analyzed using multi-faceted Rasch measurement in form of the computer program Facets (Linacre, 2006). Facets is a generalization of Wright and Master's (1982) partial credit model that makes possible the analysis of data from assessments that have more than the traditional two facets associated with multiple-choice tests (i.e. items and examinees). In the many-facet Rasch model, each facet of the assessment situation (e.g. candidates, raters, trait) is represented by one parameter. The advantages of using multi-faceted Rasch measurement is that it models all facets in the analysis onto a common logit scale, which is an interval scale. Because of this, it becomes possible to establish not only the relative difficulty of items, ability of candidates and severity of raters as well as the scale step difficulty, but also how large these differences are. Multi-faceted Rasch measurement is particularly useful in rating scale validation as it provides a number of useful measures such as rating scale discrimination, rater agreement and severity statistics and information with respect to the functioning of the different band levels in a scale.

To make the multi-faceted Rasch analysis used in this study more powerful, a fully crossed design was chosen. That is, all ten raters rated the same one hundred writing scripts on both occasions. Although such a fully crossed design is not necessary for FACETS to run the analysis, it makes the analysis more stable and therefore better conclusions can be drawn from the results (Myford & Wolfe, 2003).

Several hypotheses were formulated to guide the Rasch analysis. The first hypothesis was that a more discriminating rating scale can be seen as superior. It is important for an assessment to be able to differentiate between candidates. In the case of performance assessment, the tool that is used to achieve this is the rating scale. The more levels of candidate ability a group of raters can discern with the help of a rating scale, the better the scale is functioning. The candidate separation ratio is an excellent indicator of the discrimination of the rating scale. The higher the separation ratio, the more discriminating the rating scale is.

The next hypothesis made was that a well functioning rating scale would result in small differences between raters in terms of their leniency and harshness as a group. If a scale is functioning well, the raters will be able to discern the ability of a candidate easily and do this in agreement with other raters. Thus, raters will not vary greatly in terms of leniency and harshness. For this reason, a rating scale resulting in a smaller rater separation ratio is seen to be superior. The higher the rater separation ratio, the more the raters differed in terms of severity in their ratings.

The third hypothesis was that a necessary condition for validity of a rating scale is rater reliability (Davies & Elder, 2005). A scale that results in higher levels of rater reliability can be seen as superior. FACETS provides two measures of rater reliability: (a) the rater point-biserial correlation index (or single rater - rest of raters correlation), which is a measure of how similarly the raters are ranking the candidates, and (b) the percentage of exact rater agreement, which indicates the percentage of how many times raters awarded exactly the same score as another

rater in the sample. Both types of rater reliability statistics were deemed necessary based on Stemler's (2004) paper, in which he cautions against the use of just one statistic to describe inter-rater reliability. Recent versions of FACETS (Linacre, 2006) also include a function to calculate the percentage of exact agreement. This shows the percentage of times each rater awarded exactly the same score as another rater. The figure reported in the tables in the following section indicates the average percentage of exact agreement for all ten raters in the group.

Because rating behaviour is a direct result of using a rating scale, it was further contended that a better functioning rating scale would result in fewer raters rating either inconsistently or overly consistently (by overusing the central categories of the rating scale). The idea behind this was that if a rater is unsure what level to award when using a rating scale, the rater might either rate inconsistently or resort to a play-it-safe method and overuse the inner categories of a rating scale and avoid the outside band levels.

The measure indicating variability in raters' scores is the rater infit mean square value. Rater infit means square values have an expected value of 1 and can range from 0 to infinity. The closer the calculated value is to 1, the closer the rater's ratings are to the expected ratings. Infit mean square values significantly higher than 1.3 (following McNamara, 1996 and Myford and Wolfe, 2000) denote ratings that are further away from the expected ratings than the model predicts. This is a sign that the rater in question is rating inconsistently and therefore showing too much variation. This is called 'misfit'. Similarly, values lower than .7 indicate that the observed ratings are closer to the expected ratings than the Rasch model predicts. This is called 'overfit'. This could indicate that a rater is rating very consistently. However, it is more likely that the rater concerned is overusing certain categories of the rating scale, normally the inside values.

## 5. Results:

**1) What are the features of reader/writer interaction displayed at different levels of writing?**

Markers of reader-writer interaction were grouped into four categories: markers of writer identity, hedges, boosters and attempted passive voice.

The overall distribution of markers of ***writer identity*** can be seen in the histogram in Figure 2 below. The figure shows a heavily positively skewed distribution. Over half of the scripts made use of no markers of writer identity. However, although so many scripts did not make use of this category, the mean was just under 2.5 markers per script. This shows that a number of writers used a large number of these markers. This can also be seen in the large standard deviation of the overall sample.

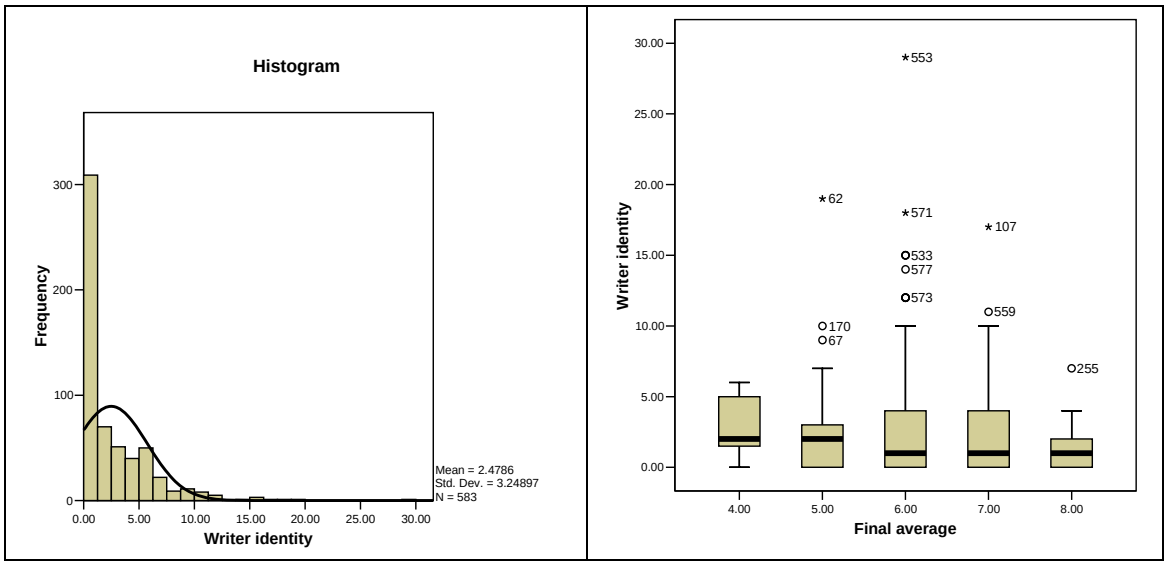

**Figure 2: Distribution of features of writer identity over overall sample and DELNA sublevels**

**Table 2: Descriptive statistics - Writer identity**

| DELNA level | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| 4 | 2.83 | 1.95 | 0 | 6 |
| 5 | 2.32 | 2.72 | 0 | 19 |
| 6 | 2.64 | 3.71 | 0 | 29 |
| 7 | 2.49 | 3.01 | 0 | 17 |
| 8 | 2.48 | 3.25 | 0 | 7 |

The box plots in Figure 2 and the table of descriptive statistics (Table 2) above show the distribution over the different DELNA levels. It is clear that this variable did not differentiate distinctly between the different proficiency levels.

The analysis of variance showed no statistically significant difference between the different levels of writing, $F(4, 577) = 1.07$, $p = .368$.

The second variable under investigation was the ***number of hedging devices***. The histogram of the overall distribution of hedges indicates a slightly positively skewed and peaked distribution. In this case, there were fewer writers who did not employ any of these devices. On average, writers used just under six of these structures per script.

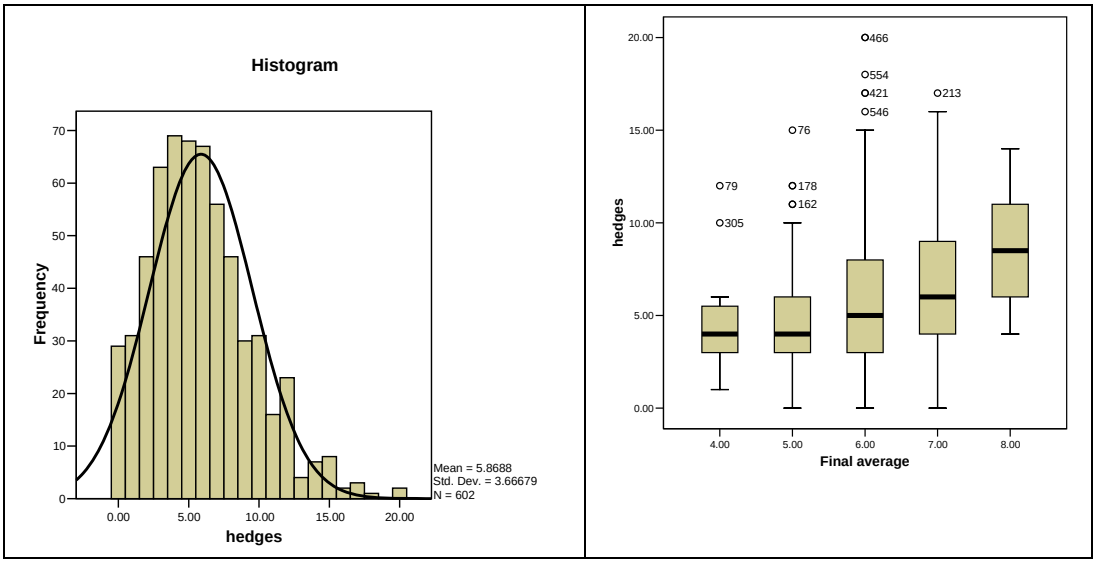


**Figure 3: Distribution of hedging devices over overall sample and DELNA sublevels**

**Table 3: Descriptive statistics – Hedges**

| DELNA level | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| 4 | 5.00 | 3.10 | 1 | 12 |
| 5 | 4.70 | 2.85 | 0 | 15 |
| 6 | 5.84 | 3.88 | 0 | 20 |
| 7 | 6.38 | 3.68 | 0 | 17 |
| 8 | 8.42 | 2.91 | 4 | 14 |

When broken up into the different DELNA band levels, the use of hedging devices can be seen to have quite clearly distinguished between different levels of writing. This is revealed in the box plot (Figure 3 above) as well as in the table summarising the descriptive statistics (Table

3 above). The table shows that whilst writers at lower levels used on average about five hedging devices in their writing, higher level writers used more than eight of these devices.

The analysis of variance revealed a statistically significant difference between the groups, F (4, 596) = 7.39, p = .000. The Games-Howell procedure showed that levels 5 and 6 as well as levels 7 and 8 were statistically distinct from each other.

The hedging variable was also investigated when script length was controlled. This showed an even stronger difference between the different proficiency levels.

The final variable investigated in this category was ***boosters***. Again, the histogram is presented below (Figure 4). This variable also had a positively skewed distribution, indicating that most writers used few of these devices; however some writers used more than ten within their piece of writing. The distribution per band level (box plots in Figure 4) and the table indicating descriptive statistics (Table 4) show that this category failed to distinguish between different levels of writing because writers of all levels used on average about 2.5 boosters in their writing.

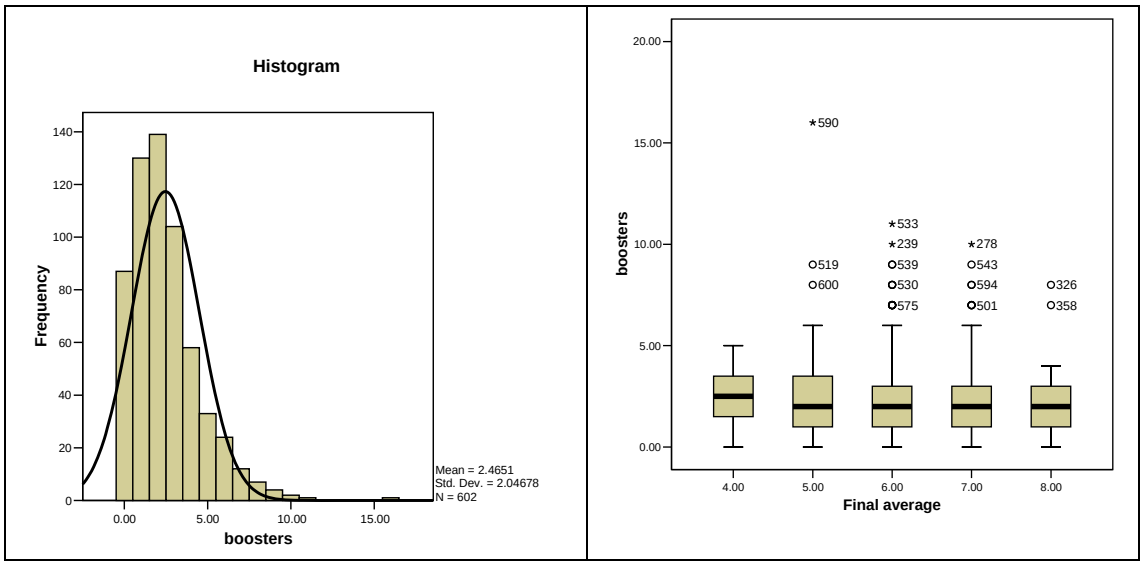


**Figure 4: Distribution of boosters over overall sample and DELNA sublevels**

**Table 4: Descriptive statistics - Boosters**

| DELNA level | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| 4 | 2.5 | 1.45 | 0 | 9 |
| 5 | 2.5 | 2.14 | 0 | 16 |
| 6 | 2.4 | 2.05 | 0 | 12 |
| 7 | 2.44 | 2.06 | 0 | 11 |
| 8 | 2.46 | 1.88 | 0 | 14 |

An analysis of variance revealed no statistically significant differences between the different levels of writing, F (4, 596) = .157, p = .960.

Finally, the use of the ***attempted passive voice*** was investigated.

Inter-rater reliability was established by a Pearson correlation between the coding of two raters on a sample of fifty scripts. The correlation coefficient shows a strong relationship, r = .898, n = 50, p = .000.

The histogram (Figure 5 below) shows that overall the passive was not attempted very frequently, with almost half the scripts not using this structure, resulting in a heavily positively skewed and peaked distribution. Other writers, however, made use of this construction up to six times within their essay. On average, the passive was used less than once per essay.

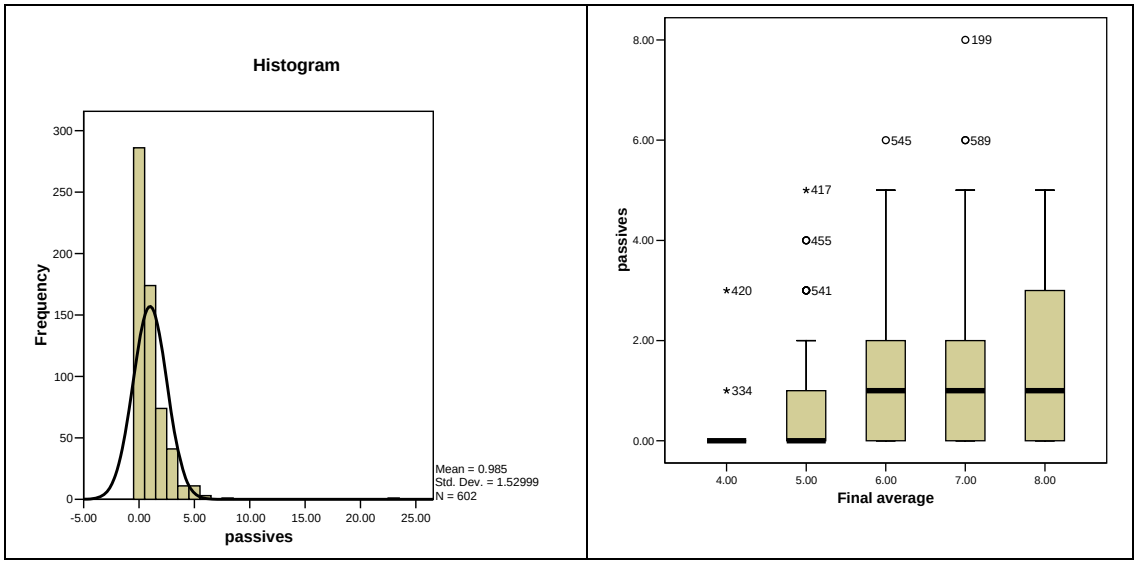


**Figure 5: Distribution of passives over overall sample and DELNA sublevels**

**Table 5: Descriptive statistics – Passives**

| DELNA level | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| 4 | .33 | .89 | 0 | 3 |
| 5 | .80 | 1.09 | 0 | 5 |
| 6 | 1.03 | 1.25 | 0 | 6 |
| 7 | 1.05 | 1.25 | 0 | 8 |
| 8 | 1.38 | 1.70 | 0 | 5 |

The box plots (Figure 5) and the table above (Table 5) show that higher level writers used the passive more frequently, whilst hardly any writers at level 4 attempted this structure; however the differences between the different levels of writing proficiency were very small on average.

An analysis of variance revealed no statistically significant differences between the different levels of writing, $F(4, 596) = 2.37$, $p = .052$

Finally, it was of interest whether there was a relationship between the use of markers of writer identity and the passive voice. It is conceivable, for instance, that writers who use markers of writer identity (by projecting their own voice into the text) use fewer passives. A correlation analysis was conducted which showed a positive relationship between these two variables, $r = .304$, $n = 583$, $p = .000$. This means that writers who used more passives also tended to use more markers of writer identity.

After this analysis, it was decided that the only measure of reader-writer interaction that could be transferred into the rating scale was the measure of hedging. The rating scale can be seen in Table 6 below.

**Table 6: Rating scale – Reader-writer interaction**

| 9 | 8 | 7 | 6 | 5 | 4 |
|---|---|---|---|---|---|
| More than 9 hedging devices | 7-8 hedging devices | 5-6 hedging devices | 3-4 hedging devices | 1-2 hedging devices | No hedging devices |

The analysis showed distinct levels in the number of hedging devices. The descriptors were scaled to match the levels in the DELNA scale and to allow for clear differentiation between levels.

**2) How reliable and valid is this trait scale for hedging when compared to the previously existing scale for academic style?**

The Rasch analysis set out to compare the two trait scales for academic style. The existing scale had descriptors pertaining to academic style in general, whilst the new scale had descriptors only for the use of hedging.

Table 7 below displays the rating scale statistics for the two trait scales. The new scale for hedging was clearly more discriminating (in this case both scales have six levels) with a candidate separation ratio of 5.86 compared to the separation ratio of 3.32 for the existing scale. This means that the raters were able to distinguish more levels of candidate ability when using the new scale.

**Table 7: Rating scale statistics for style**

| DELNA scale - style | New scale - hedging |
|---|---|
| Candidate discrimination:<br>Candidate separation ratio: 3.32 | Candidate discrimination:<br>Candidate separation ratio: 5.86 |
| Rater separation and reliability:<br>Rater separation ratio: 5.56<br>Rater point biserial: .78<br>Exact agreement: 37.2% | Rater separation and reliability:<br>Rater separation ratio: 1.94<br>Rater point biserial: .89<br>Exact agreement: 53.7% |
| Variation in ratings:<br>% Raters infit high: 20%<br>% Rater infit low: 20% | Variation in ratings:<br>% Raters infit high: 10%<br>% Rater infit low: 10% |

The raters rated considerably more similarly in terms of severity when using the new scale. Furthermore, both the inter-rater reliability statistics were significantly higher than those of the existing scale. Fewer raters rated with too much or too little variation (only 20% of the raters compared to 40% of raters when applying the existing scale). A closer scrutiny of the use of the different band levels showed that the raters (as a group) displayed a strong central tendency effect when using the existing scale – levels 4, 5 and 9 were underutilized.

**3) What are rater's perceptions of using the two rating scales?**

To establish the raters' perceptions of the efficacy of the two scales, first a questionnaire was administered, and then a subset of seven raters were interviewed.

The questionnaire questions focused only on the new scale. Raters were asked to comment on what they thought of the trait scale for hedging. A summary of the results with illustrative comments can be seen in Table 8 below.

**Table 8: Summary of questionnaire responses**

| **Trait scale – Hedging devices** |
|---|
| Positive comments: 6 raters<br>• '*this is a great addition as the previous scale lacked such subtlety*' (Rater 7)<br>• '*good, interesting, effective*' (Rater 8) and Rater 1 noted that '*the scale was easy to follow*' |
| Negative comments: 4 raters<br>• '*some good scripts managed with no hedging. Is it really necessary? Does it show lack of understanding of academic style to do without? Probably, yes. Many picked up hedging from the question. That should have given all the hint that it was necessary to remember to use it*' (Rater 4)<br>• '*I was not quite satisfied with this category as I felt that it only measured explicit hedging devices*' (Rater 10) |

So whilst more raters commented positively about the trait scale for hedging, there were some criticisms voiced. Rater 4 suggested that hedging was not the only manifestation of academic style in writing, which is of course a very valid comment. The same idea is reflected in Rater 10's comment from an interview below:

> Rater 10: I just wasn't sure with, I guess the hedging devices would be an example, mmh, sometimes I might think it was actually a pretty good script, but they just hadn't put any hedging devices in and so I felt like I was marking them down for something that they didn't know they were supposed to do. And that they could maybe produce a pretty good piece of work without having hedging devices and no kind of account was taken. So I guess this [the new scale] seemed a bit more rigid to me and maybe not fitting each individual case.

Although the analysis presented under research question 1 was not able to establish more categories of academic style suitable for inclusion in a rating scale, this does not mean that other measures should not be pursued in the future.

Raters further criticized the fact that some information was lost because the descriptors in the new scale were too specific. Rater 5, for example, argued that a simple count of hedging devices could not capture variety and appropriateness:

> Researcher: You said that, other than hedging, style wasn't really considered.
> Rater 5: Yeah, it does seem a bit limited. And then they might repeat the same hedge and they might copy the one from the prompt and so they get automatic points which I suppose is a strategy you can use when you are doing academic writing, but quite often sort of non-native speakers will rely on one or two hedges all the way through [...] Whereas the good writers will very sparingly use hedges but they will use them just right and they will vary them. [...] So maybe something about variety of hedges and appropriateness as well. I suppose that is similar [to the DELNA scale] it sort of relies on the marker's knowledge of English in a more kind of global way sort of. But maybe that is the inter-rater reliability issue coming up. So the DELNA scale allows me the flexibility to use my own judgement about a script in all categories.

The positive comments reported in Table 1 above, were also reflected in the interviews. The idea of being able to arrive precisely at a score was seen as a great advantage of the scale for hedging devices. This can be seen in the comment by Rater 7 below.

> Rater 7: It is interesting, I found that it [the new scale] is quite different to the DELNA one and it is quite amazing to be able to count things and say, I know exactly which score to use now.

One of the most unexpected themes emerging from the interviews was the fact that almost all raters reported a ***changed rating behaviour since using the new scale***. The first rater interviewed (Rater 3) raised the topic and it was then included in the interviews that followed. Here is what Rater 3 said:

> Rater 3: Yeah. I found the first time round, there was definitely an improvement in my DELNA marking
> Researcher: In what way?
> Rater 3: It made me more aware, I hadn't really thought about hedging very much, I have to say, mmh, so that then I started to notice them, so there is, it has had a very positive spin-off. It has pinpointed things, because the DELNA one is less specific, it is less specific, so this, the two kind of go together quite nicely, this [the DELNA scale] pinpoints things. But by marking with the new scale, it has, I have got in my mind now, I can see hedging
> Researcher: So maybe like a training scale?
> Rater 3: Yeah, it definitely has been very useful. It is sort of more awareness of things which I might have glossed over [...]

This very interesting idea of the scale being useful as a training tool will be discussed below.

## 6. Discussion

Four different categories of reader/writer interaction were analysed to answer Research question 1: hedges, boosters, writer identity and attempted passive voice. The analysis of the writing scripts showed that lower level writers used fewer ***hedges*** than writers of higher proficiency. This in fact was by far the most discriminating measure identified in this category. Very little research has looked at differences over proficiency

levels in the use of hedging. Most prior research has focussed on the features of writing found in particular groups of writers (e.g. Chinese L2 writers – Hu, Brown and Brown, 1982 or EFL writers – Bloor and Bloor, 1991) or a comparison between groups of writers (e.g. L1 and L2 students – Hyland and Milton, 1997). Intaraprawat and Steffensen (1995), however, were able to show that better L2 writers used twice as many hedges as poor writers. The same finding was reported by Kennedy and Thorp (2002) in the context of the IELTS writing task. All these findings, as well as the results of the current study, suggest that including the category of hedges into the rating scale descriptors is warranted, a practice not common in current scales. Further research on this topic is desirable.

Another category of reader/writer interaction investigated in both the pilot study and the main analysis, was ***boosters***. A number of studies (Allison, 1995; Bloor & Bloor, 1991; Hyland & Milton, 1997; Kennedy & Thorp, 2002) showed that L2 writers (especially at lower levels) overuse boosters in their writing. Similarly, Intaraprawat and Steffenson (1995) were able to show that lower level ESL writers use more than double the number of boosters as higher level ESL writers. This study, with a mixed cohort of L2 and L1 writers, was not able to show that lower level writers used more boosters than higher level writers, as might be hypothesised based on the findings of previous research. It might be necessary to do a more fine-grained, qualitative analysis of the type of boosters used to identify differences between writers of higher and lower proficiency.

The third category of reader/writer interaction investigated was that of ***writer identity***. The prior research findings on this topic were mixed. Hyland (2002a; 2002c) showed that L2 writers use fewer personal pronouns than L1 writers and that this inevitably resulted in a loss of voice. On the other hand, Shaw and Liu (1998) were able to show that as L2 students develop their writing, they slowly move away from the use of personal pronouns and toward using passive verbs. This study showed that the students investigated in this context used very few

expressions of writer identity. There were furthermore no significant differences between the writers at different levels. It is possible that the particular genre (expository writing) investigated in this study does not lend itself to the expression of writer identity.

The fourth category of reader/writer interaction investigated was the use of the ***passive voice***, a category related to writer identity. In line with what Shaw and Liu (1998) and Banerjee and Franceschina (2006) found, it was seen that as the writing proficiency level increased, more instances of passive voice were found. However, overall the frequency was very low. This again might be a feature of the genre of the task. There was furthermore no negative correlation between the use of passives and instances of writer identity, as might be expected. The interaction between these two devices clearly warrants further research.

Research questions 2 and 3 suggested that the category of hedging performed well when the quantitative data were analyzed, outperforming the existing rating scale of style in all aspects. The raters' comments in the questionnaire were also generally positive, although some raters thought that a script could be highly successful without hedging devices. It is clear that the category of hedging provides a substantially narrower picture of a writers' academic style than its broader counterpart in the DELNA scale. The vaguer descriptors in the DELNA scale, however, resulted in a central tendency effect. Hardly any raters used the outside scale categories. This was possibly the case because raters did not know what specific features to focus on. In Phase 1 of this study, several aspects of style were pursued, but the only one that successfully discriminated between the levels was the category of hedging. The category of hedging, although functioning well, is clearly just one aspect of academic style. Future revisions of the scale will hopefully include a wider variety of features of academic style. For example, it might be interesting to investigate whether the category of voice used by Cumming et al. (2005) is a meaningful measure for the type of writing genre investigated in this study. Another feature not generally

seen in academic writing might be the use of contractions or certain lexical items which are considered as colloquial.

Finally, it is also important to consider practicality, an important consideration in test design. The new scale is clearly more laborious to design and therefore not as practical for classroom teachers to develop. In terms of scale use, however, there seemed to significant difference in the time raters needed when applying the two scales.

## 7. Conclusion

The study presented has a number of implications. The first relates to rating scale development and the level of description in rating scale level descriptors. Two opposing views can be voiced. Firstly, there are proponents of high levels of descriptions in the scale descriptors so that raters are provided with a maximum amount of guidance. This approach, as was shown with the new scale developed in this study, has the advantage of achieving high levels of rater reliability, but, has the downside that certain aspects of raters' knowledge and experience will not be tapped into and are therefore wasted. This could result in construct under-representation. This was to some extend argued by raters in the qualitative part of this study. A case can also be made for the opposing view. The pre-existing trait scale for academic style was marked by very vague, impressionistic rating scale descriptors. This provides a wider window for rater interpretation of the meaning of the descriptors, but, as was shown in this study, inevitably results in lower inter-rater reliability. The most alarming finding of this study was, however, that raters almost exclusively relied on the inner band levels of the rating scale when grading the writing scripts, which were chosen to represent a wide range of levels. It can therefore be argued, that the more impressionistic terminology of the rating scale, whilst leaving more room for raters to use their own knowledge, also lacks construct validity, as raters were only able to identify few measurable differences between the

writing scripts in terms of academic style. There is probably no immediate solution to resolve this tension between the two opposing arguments. What is clear is that further research into the construct of academic style is necessary and that language testing theory needs to continue to be informed by findings from areas like discourse analysis and second language acquisition (e.g. Bachman & Cohen, 1998). Once further advances in these areas broaden our understanding of the construct of 'academic style', this should directly inform theory and practice in language testing, as was demonstrated in this study.

Another implication for rating scale development is, that it is important to ascertain if features deemed to discriminate between learners' writing ability in fact do so in practice. It would, for example, be a mistake to ask raters to rate down scripts making repeated use of markers of writer identity, when this is in fact a feature of writing of all proficiency levels and therefore the measure is not able to discriminate between scripts at a number of levels. Similarly, raters should be made aware during rater training which features are commonly displayed at certain levels of writing.

The final implication relates to rater training. A number of raters in the study noted that using the more detailed, empirically-developed rating scale, resulted in subsequent changed rating behavior. This was acknowledged to be a useful side-effect of this research project. Raters reported an awareness-raising effect from using the rating scale for hedging. It could therefore be argued that this level of detail, even if not included in an operational scale, might be productive in training scales used for new researchers or during training sessions to raise the raters' levels of alertness to certain features of discourse. Especially newer, less experienced raters should be provided with such tools when first joining a rating program.

## 8. References:

Allison, D. (1995). Assertions and alternatives: Helping ESL undergraduates extend their choice in academic writing. *Journal of Second Language Writing, 4*, 1-16.

Bachman, L. F., & Cohen, A. D. (Eds.). (1998). *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Banerjee, J., & Franceschina, F. (2006). *Documenting features of written language production typical at different IELTS band score levels*. Paper presented at the Workshop sponsored by the European Science Foundation entitled 'Bridging the gap between research on second language acquisition and research on language testing', Amsterdam, February 2006.

Barlow, M. (2002). *MonoConc Pro 2.2*. Houston: Athelstan.

Bloor, M., & Bloor, T. (1991). Cultural expectations and socio-pragmatic failure in academic writing. In P. Adams, B. Heaton & P. Howarth (Eds.), *Academic writing in a second language: Essays on research and pedagogy*. Norwood, NJ: Ablex.

Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication, 10*, 39-71.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 1-75.

Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum.

Elder, C. (2003). The DELNA initiative at the University of Auckland. *TESOLANZ Newsletter, 12*(1), 15-16.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online rater training program. *Language Testing, 24*(1), 37-64.

Elder, C., & Erlam, R. (2001). *Development and validation of the diagnostic English language needs assessment (DELNA): Final Report*. Auckland: University of Auckland, Department of Applied Lnaguage Studies and Linguistics.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175-196.

Elder, C., & von Randow, J. (2002). *Report on the 2002 Pilot of DELNA at the University of Auckland*. Auckland: University of Auckland, Department of Applied Language Studies and Linguistics.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal, 41*(4), 287-291.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Hu, Z., Brown, D., & Brown, L. (1982). Some linguistic differences in the written English of Chinese and Australian students. *Language Learning and Communication, 1*(1), 39-49.

Hyland, K. (1997). Qualification and Certainty in L1 and L2 Students' Writing. *Journal of Second Language Writing, 6*(2), 183-205.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Hyland, K. (2000a). Hedges, boosters and lexical invisibility: Noticing modifiers in academic texts. *Language Awareness, 9*(4), 179-301.

Hyland, K. (2000b). 'It might be suggested that...': Academic hedging and student writing. *Australian Review of Applied Linguistics, 16*, 83-97.

Hyland, K. (2002a). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics, 34*, 1091-1112.

Hyland, K. (2002b). Directives: Argument and Engagement in Academic Writing. *Applied Linguistics, 23*(2), 215-239.

Hyland, K. (2002c). Options of identity in academic writing. *ELT Journal, 56*(4), 351-358.

Hyland, K., & Milton, J. (1997). Hedging in L1 and L2 student writing. *Journal of Second Language Writing, 6*(2), 183-296.

Intaraprawat, P., & Steffensen, M. S. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing, 4*(3), 253-272.

Kennedy, C., & Thorp, D. (2002). *A corpus-based investigation of linguistic responses to an IELTS academic writing task*: University of Birmingham.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training raters online: How does it compare with face-to-face training? *Assessing Writing*.

Linacre, J. M. (2006). Facets Rasch measurement computer program. Chicago: Winsteps.

McNamara, T. (1996). *Measuring second language performance*. Harlow, Essex: Pearson Education.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system*. Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. TOEFL Monograph 24. Princeton: Educational Testing Service.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

Shaw, P., & Liu, E. T.-K. (1998). What develops in the development of second-language writing. *Applied Linguistics, 19*, 225-254.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation, 9*(4), 1-20.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111.

Vande Kopple, W. J. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36*, 82-93.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

**Appendix 1: List of hedges, boosters and markers of writer identity analysed during Phase 1**

| **Writer Identity** |
|---|
| I, you, we, us, our, me, mine, yours, my, your |
| **Hedges** |
| Can, could, may, might, perhaps, maybe, possible/possibly, suppose/supposed, I think, I feel, sometimes, seem, relative/relatively, would, appear, probably, possibility, fairly, usually, tend, hardly, more or less, should, suggest, indicate, potential/ly, assume, generally, about, believe, hypothesise, likely, speculate, estimate, doubt (used without a negative), presume |
| **Boosters** |
| Certain/ly, clear/ly, I know, definite/ly, fact, obvious/ly, sure/ly, like/ly, significant/ly, enormous/ly, no/never, a lot, really, main/ly, very, extremely, at last, major, always, demonstrate, substantially, will, all, many, apparent, evident, doubt (used in negative sense, i.e. no doubt), doubtless, indeed, of course |